Available online at www.sciencedirect.com
ScienceDirect
Procedia Engineering 123 (2015) 291 – 299
Procedia Engineering
www.elsevier.com/locate/procedia

Creative Construction Conference 2015 (CCC2015)

Establishing formalized representation of standards for construction cost estimation by using ontology learning

Zhe Liu[a], Zhiliang Ma[a]*

[a]*Tsinghua University, Haidian District, Beijing, 100084, China*

Abstract

Construction cost estimation for tendering is important for both tenders and bidders in construction projects and needs to be strictly complied with corresponding standards so that quantities and prices from different bidders are comparable. In the view of flexibility and extensibility, ontology is regarded as a promising technology for formalized representation of the standards for construction cost estimation (cost standards for short hereafter) in computer programs. In order to automate the processes of construction cost estimation for estimators. However, the manual establishment of ontology for construction cost estimation (cost ontology for short hereafter) is labor-intensive and time-consuming for software developers, not to mention that there are numbers of standards for different types of construction projects in different regions. In order to solve this problem, a semi-automatic approach based on the framework of cost ontology that authors established previously is proposed to establish the cost ontology by using ontology learning technology. Firstly, the data sources, i.e. the cost standards are analyzed and the corresponding relations between information in cost standards and the elements in the framework are summarized. Then based on the corresponding relations, the approach is designed, in which concepts, relations and rules are extracted by natural language processing and domain lexical analysis to fill the framework. The approach lays a foundation for the practical use of ontology for automating construction cost estimation.

Peer-review under responsibility of the organizing committee of the Creative Construction Conference 2015

Keywords: construction cost estimation; formalized representation; ontology learning

1. Introduction

Construction cost estimation for tendering is important for both tenders and bidders in construction projects and needs to be strictly complied with corresponding standards so that quantities and prices from different bidders are

* Corresponding author. Tel.: +86-10-6277-3543
E-mail address: mazl@tsinghua.edu.cn

Peer-review under responsibility of the organizing committee of the Creative Construction Conference 2015
doi:10.1016/j.proeng.2015.10.093

comparable. In order to improve the efficiency and accuracy of the estimation, standards for construction cost estimation (cost standards for short hereafter) are manually translated into formalized representation such as hard coding, parametric table, decision tree, etc. in computer programs to semi-automate the processes [1]. However, it is difficult to extend, share and reuse these formalized representations for other applications. In recent years, ontology has been regarded as a promising technology to represent concepts, relations and rules for both human and program on account of its flexibility and extensibility. Additionally, it can be used for automatic decision processes with the support of existing reasoning machines and has been adopted in researches on knowledge representation [2,3,4], information query [5,6], rule checking [7,8], etc. in the AEC (Architecture, Engineering, and Construction) industry.

In the domain of construction cost estimation, Staub-French et al. [9] proposed a feature-based ontology for estimators to generate and maintain complete, consistent and expeditious estimates. But limited to the ontology technology at that time, this ontology was not sharable. Lee et al [10] suggested a BIM (Building Information Modeling) and ontology-based approach for construction cost estimation which utilized reasoning machines to infer work items and unit costs automatically. The authors [11] made a preliminary discussion on adopting ontology as a new formalized representation of cost standards, and [12] established a framework of ontology for construction cost estimation (cost ontology for short hereafter) to discriminate building products into corresponding cost items. Although cost ontologies are sharable and reusable and can be utilized to improve the efficiency and accuracy of construction cost estimation for estimators, the manual establishment of ontology is labor-intensive and time-consuming for software developers. Since there are numbers of standards for different types of construction projects (e.g. civil buildings and industrial buildings) in different regions (e.g. provinces in China), cost ontologies are still hampered from practical use.

Ontology learning is a technology to create ontology semi-automatically based on techniques such as natural language processing, machine learning, data mining, etc., and has been implemented in web search successfully [14]. According to the types of data sources, ontology learning can be classified into three groups, i.e., ontology learning from structured data (e.g. database schemas), that from semi-structured data (e.g. web pages), and that from unstructured data (e.g. text documents). For each type of data sources, there have been several existing methods and tools to extract concepts and relations to create ontologies [13,14]. A few researches have focused on ontology learning from data sources in Chinese [15,16,17]. Nevertheless, for the full extraction of information from the semi-structured Chinese cost standards to establish cost ontologies, the existing ontology learning tools are still insufficient.

In order to solve this problem, a semi-automatic approach based on the framework that authors established previously is proposed to establishing the cost ontology by using ontology learning technology. Firstly, the data sources, i.e. the cost standards are analyzed and the corresponding relations between information in cost standards and the elements in the framework of cost ontology are summarized. Then based on the corresponding relations, the approach is designed, in which concepts, relations and rules are extracted by using natural language processing and domain lexical analysis to fill the framework of cost ontology. Finally, the approach is implemented and tested with typical Chinese cost standards.

2. Analysis of two typical cost standards

2.1. Cost standards for construction cost estimation

The bill-of-quantity (BQ for short hereafter) method is well-accepted for tendering in practice in many countries and regions all over the world. According to the BQ method, buildings are broken into building products and classified into groups with similar features and construction works. For each construction work, several construction conditions are specified with different unit costs of labor, material, and equipment etc., i.e. quota items. In each BQ item, once all related quota items are identified, the comprehensive unit cost of BQ item can be calculated, and the total quantities of building products are classified into the same BQ item can be summed up. Then the budget cost of buildings is obtained by summarizing the product of the comprehensive unit cost and the total quantity of each BQ item.

The BQ items and quota items are specified in BQ standards and quota compositions separately. The breakdown structure of building products and construction works differ between different types of construction projects while the unit costs vary from region to region. Thus there are numbers of BQ standards and quota compositions in China. It is noted that, in China, few construction firms record history data of construction and summary their own quota compositions. They follow the quota compositions published by the government and adjust consumptions and prices if necessary. In this way, the quota compositions in China act as cost standards, too.

In this paper, two typical cost standards are analyzed, i.e. the BQ standard "Standard method of measurement for building construction and fitting-out works GB 50854-2013" (GB 50854 for short hereafter) and the quota composition "Quota set for building construction and fitting out works in Beijing" (Beijing Quotas for short hereafter).

2.2. Cost standards for ontology learning

Since the GB 50854 and the Beijing Quotas have been input into databases respectively, from a general viewpoint, they are structured data sources for ontology learning. Nevertheless, for a full extraction, texts in certain fields of records in the database need to be processed as unstructured data sources, and their lexical characteristics should also be analyzed. It is noted that, some information (e.g. material list for quota item) is stored in several fields in different tables in the relational database. In the GB 50854, BQ items are classified in a "domain - section - subsection" hierarchy. Table 1 shows an example of BQ items in a subsection "cast-in-place concrete column" in GB 50854. Each BQ item contains six fields, i.e. code, name, essential features, unit, rules of quantity takeoff, and construction works. Building products are mainly classified by the subsections and BQ item names.

Table 1. Example of BQ items in a subsection "cast-in-place concrete column" in the GB 50854.

Subsection: Cast-in-place concrete column					
Code	Name	Essential features	Unit	Rules of quantity takeoff	Construction works
010502001	Rectangular column	1. Type of concrete 2. Strength grade of concrete	m3	…	1. Manufacture, installation, uninstallation, storage, transportation and cleanness of concrete form 2. Pouring, vibrating, and curing of concrete
010502002	Structural column	1. Type of concrete 2. Strength grade of concrete	m3	…	Ditto
010502003	Special-shaped column	1. Cross shape of column 2. Type of concrete 3. Strength grade of concrete	m3	…	Ditto

For instance, if a building product is a column with material of cast-in-place concrete and cross profile of rectangle, it should be classified into the BQ item "rectangular column", coded as "010402001" in the subsection "cast-in-place concrete column", as shown in Fig 1. In the authors' framework of cost ontology [12], the "column", "cast-in-place", "concrete", "rectangle" are extracted as concepts used to represent the "value of discrimination feature", while the "type", "material", and "cross profile" are extracted as relations between building products and the feature values.

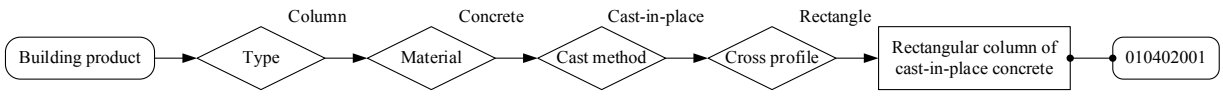

Fig. 1. Discrimination of BQ item "rectangular column", coded as "010402001" in the subsection "cast-in-place concrete column".

In the Beijing Quotas, quota items are organized in a "domain - section - subsection - construction work" hierarchy, and each quota item contains at least six fields, i.e. code, name, item, labor, material list, and equipment list, as shown in Table 2. For some quota items, extra fields are added to describe the construction methods, which are extracted as "construction features" and the corresponding values in the framework of cost ontology. According to subsection, name, construction work and extra fields, quota items are selected to correspond to BQ items similarly.

For other concepts and relations in the framework of cost ontology, the corresponding data sources are summarized and listed in Table 3.

Table 2. Example of quota items for construction work "pouring, vibrating, and curing of concrete" in the Beijing Quotas.

Subsection: cast-in-place concrete column							
Construction work: pouring, vibrating, and curing of concrete, for 1m^3 of volume of column (CNY)							
Code	Name	Item	Labor	Material		Equipment	
				C30 premixed concrete	…	Mortar mixer (200L)	…
		Unit	day	m^3	…	day	…
		Unit cost (CNY)	74.30	410.00	…	11.00	…
5-7	Rectangular column	Unit consumption	0.686	0.9860	…	0.0052	…
5-8	Structural column		1.231	0.9860	…	0.0052	…
5-9	Special-shaped column		0.971	0.9860	…	0.0052	…

Table 3. Data sources for major concepts and relations in the framework of cost ontology.

Concepts and relations in the framework of cost ontology		Data sources
Information subject (Concept)	Building product	Subsection fields and name fields of BQ items
	BQ item	Each record of BQ item
	Quota item	Each record of quota item
	Construction work	Construction works of both BQ items and quota items
	Construction material and equipment	Material list and equipment list of quota items
Value of feature (Concept)	Value of discrimination feature for BQ item	Subsection names and item names of BQ items
	Value of essential feature for BQ item	Material list, equipment list, extra fields of quota items
	Value of construction feature for construction works	Extra fields of quota items
Feature (Relation)	Discrimination feature for BQ items	N/A
	Essential feature for BQ items	Essential features of BQ items
	Construction feature for construction works	N/A

3. Approach to establishing the cost ontology using ontology learning technology

Generally, in the process of ontology learning from structural data, concepts and relations can be extracted to establish ontology right after the database schema is identified. For cost ontology, since some concepts and relations are in the same fields of records, extra steps for texts in the fields are needed. The primary ontology learning process consists of the following five steps, and the information flow is shown in Fig 2.

Step 1 Read relevant contents from the database.
Step 2 Tokenize and tag texts in the contents as basic domain terms.
Step 3 Fill terms into the framework of cost ontology to establish the hierarchy of concepts.
Step 4 Classify adjectives into different categories as values of features, and label features according to the values.
Step 5 Translate BQ items and quota items into definitions for discrimination and establish cost ontology.

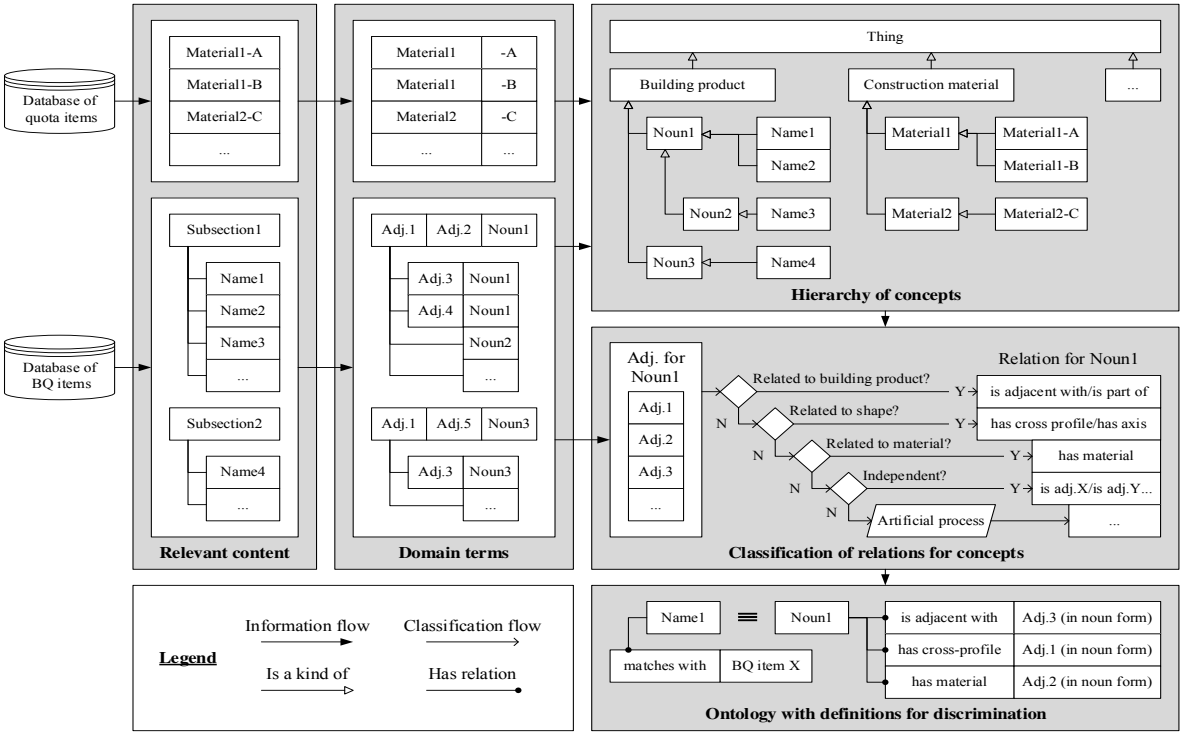

Fig 2. Illustration of information flow of ontology learning from cost standards.

In the view of output, the extraction of concepts corresponds to the first three steps, while the extraction of relations is achieved by Step 4. Rules for discrimination of BQ items and quota items are established by Step 5.

3.1. Extraction of concepts

In Step 1, texts are read from subsection fields and name fields of BQ items or quota items from databases of cost standards. In Step 2, the texts are tokenized into terms and tagged as nouns or adjectives by existing lexical analyzer for Chinese [18]. For instance, “rectangular column” are tokenized and tagged as noun “column” and adjective “rectangular”.

Based on the analysis of cost standards, naming patterns of BQ items and quota items are summarized. Since BQ items or quota items are grouped as subsections, the nouns in name fields of BQ items or quota items should be the sub-concepts of nouns in subsection fields. In addition, texts in name fields are compound words, and can be treated as sub-concepts of nouns. For example, BQ item “cushion layer” is in the subsection “cast-in-place concrete footing”, so “cushion layer” is deduced as a sub-concept of “footing”. Other contents such as material lists and equipment lists, have similar naming patterns.

In Step 3, according to the naming patterns, all nouns are filled into the framework of cost ontology as sub-concepts of building product, construction work, construction material and equipment correspondingly to establish the hierarchy of concepts.

3.2. Extraction of relations

In the authors' framework of cost ontology, adjectives act as value of features in the discrimination of BQ items and quota items. Although some general background knowledge sources [19,20] exist for querying the semantic meaning for terms, they do not contain enough terms in a specific domain, i.e. construction cost estimation, for the classification of adjectives in this case. In addition, the discrimination feature for BQ items and construction feature for construction work have no corresponding data sources, as listed in Table 3. These features should be labeled after extraction. Thus, a special method based on the hierarchy of concepts in the previous steps is designed to classify and label the features described by adjectives.

For each noun, all possible adjectives are classified by the following five rules, as shown in Fig 2.

Rule 1 If the noun form of an adjective matches certain noun of building products or parts, the adjective is describing the spatial relations, which can be labeled as "is adjacent with" or "is part of", etc. For instance, "footing" in "footing beam" means that the beam is part of the footing.

Rule 2 If an adjective matches the pattern "X shape" in Chinese, it is describing the shapes of building products. The corresponding features can be labeled as "has cross profile" or "has axis".

Rule 3 If the noun form of an adjective matches the noun of construction materials, the noun is one of the values of "has material" feature.

Rule 4 If an adjective is "irrelevant" with all other possible adjectives, it is marked as "independent", and a feature is created and labeled as "is X". For instance, "structural" in the "structural column" is independent and a feature labeled "is structural" is created for columns.

Rule 5 If an adjective does not match any rules above, it should be processed manually.

Two adjectives are defined as "irrelevant" in the following two cases. Since descriptions in cost standards are precise and concise, once two adjectives appear at the same time for the same noun and there is no explicit indicator shows that they are "or" relation, they should describe different aspect of the noun. Take the BQ item of "structural column" in subsection "cast-in-place concrete column" for instance, "structural", "cast-in-place" and "concrete" describe the function, construction method and construction method of the column separately. In the other case, if two adjectives for the same noun never appear in the same subsection, they should be describing the noun from different viewpoints, and marked as "irrelevant", too.

3.3. Translation of definitions of BQ items and quota items for discrimination

Based on concepts and relations extracted above, the definitions of BQ items and quota items for discrimination can be easily established by translation. For instance, the BQ item "010402001 - rectangular column of cast-in-place concrete" is translated into a similar form like "column, and has cross profile of some rectangle, and has material some concrete, and use cast method some cast-in-place" according to the syntax of certain ontology language. Once the design result is transformed into ontology data, the corresponding BQ items and quota items can be determined according to the definitions by reasoning machines [10].

4. Module and workflow of the semi-automatic approach

Corresponding to the five steps stated above, five modules are defined respectively, i.e. cost standard reader, lexical analyzer, naming pattern parser, feature value classifier, and cost ontology writer, as shown in Fig 3. The initial input, i.e. the cost standard database and cost ontology framework, are processed by the five modules and the cost ontology and some unclassified feature values are generated automatically. Newly generated concepts and relations in the cost ontology are checked manually. If they are generated correctly, they are marked as "confirmed".

If not, there are three kinds of exceptions that need to be corrected manually.

Exception 1 The texts in the cost standard database contain the domain terms that the lexical analyzer cannot tokenize and tag correctly. It needs to be corrected by adding the domain terms in the customized dictionary for the lexical analyzer.

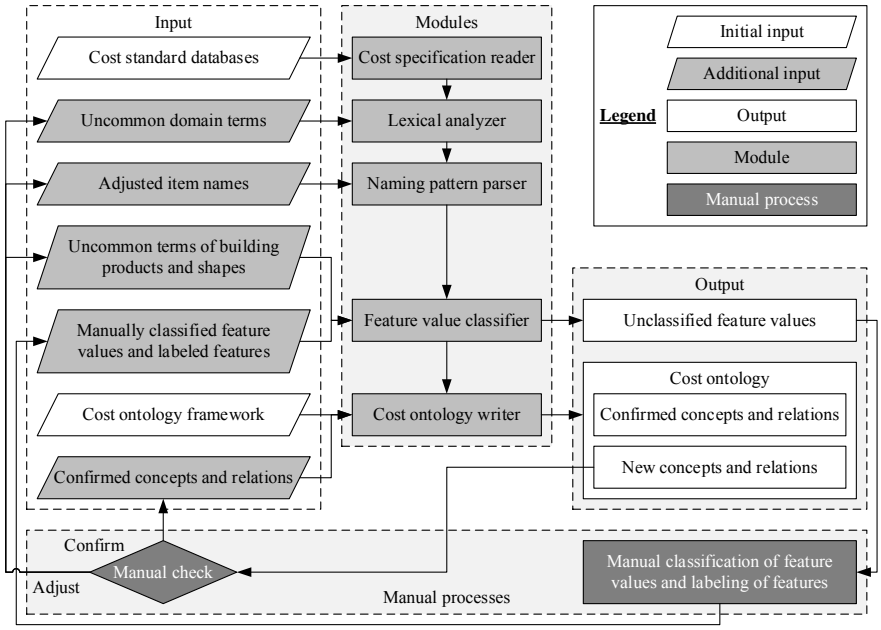

Fig 3. Module and workflow design of the semi-automatic approach.

Exception 2 There are minor special item names that do not match the naming patterns of cost standards so that the naming pattern parser cannot recognize them correctly. It needs to be corrected by adding records of synonyms that match the naming patterns to replace the original item names in the naming pattern parser.

Exception 3 There are minor special feature values that are classified into wrong features. It needs to be corrected by adding records of semantic tags to mark these feature values, i.e. whether they are relevant to the building products or cross shapes.

The unclassified feature values need to be classified and the corresponding features need to be labeled manually. All manual adjustments, classifications, and labeling are recorded. Along with the initial input, the additional records are processed by the five modules again. The generated cost ontologies are improved by iterations until they are ready for practical use.

5. Implementation and verification

The modules are implemented in Python. At present, it has no graphic user interface. All additional input and unclassified feature values are written in text files while the generated cost ontologies are exported as OWL (Ontology Web Language) files which can be opened in a general-purpose ontology editor, such as Protégé [21], as shown in Fig 4.

The effectiveness of the semi-automatic approach is estimated by the percentage of manual operations including adjustment, classification and labeling, etc. The approach has been applied to the section “concrete construction of

reinforced concrete building" of GB 50854 and the Beijing Quotas for verification. Products are defined with features and values, and linked with proper BQ items and quota items, which can be used for reasoning.

Totally, 156 manual operations were required for 230 items in the reinforced concrete building section. Then 877 concepts and 2917 relations were output in the cost ontology. It actually improved the efficiency in the establishment of cost ontology. For each manual operation, detailed statistic data is listed in Table 4. It is noted that, the terms of shapes need highest percentage of manual operations because there is no explicit data source of shapes for the modules to learn from. Nevertheless, these shapes can be reused in other sections, thus the percentage of manual operations will decrease in a larger verification of more sections.

Fig 4. Cost ontology established semi-automatically (opened in Protégé with some key information translated into English).

Table 4. Statistic data for verification of the semi-automatic approach.

	Manual operations	Total input or output	Percentage of manual operations
Domain terms	38	877	4
Item names	31	230	13
Terms of buildings	34	336	10
Terms of shapes	16	37	43
Feature values	37	121	31

6. Conclusion and future works

In this paper, a semi-automatic approach based on the framework that authors established previously was proposed to establish the cost ontology by using ontology learning technology. Firstly, the data sources, i.e. the cost standards, were analyzed and the corresponding relations between information in cost standards and the elements in the framework of cost ontology were summarized. Then based on the corresponding relations, the approach was designed, in which concepts, relations and rules were extracted by using natural language processing and domain lexical analysis to fill the framework. Finally, the approach was implemented and tested with typical Chinese cost standards. The approach lays a foundation for practical use of ontology for automating construction cost estimation.

The biggest obstacle to fully automatic cost estimation by using ontology is that some information required by the reasoning based on cost ontology cannot be found in the design model, such as design data based on IFC (Industry Foundation Classes). This problem is caused by both technological factors and legal factors as discussed in the authors' previous study [22]. In the future, the research will focus on the mapping between IFC schema and the cost ontology established from a full set of sections. During the mapping, the missing information for cost ontology in the IFC schema can be analyzed in depth. After that, an effective mechanism to fill the missing information should be proposed and implemented.

Acknowledgements

This research is supported by "National Natural Science Foundation of China" (No. 51278279), "Tsinghua University Research Fund" (No. 2011THZ03) and Glodon Software Company Limited.

References

[1] Ma, Z., Wei, Z., Zhang, X. Semi-automatic and specification-compliant cost estimation for tendering of building projects based on IFC data of design model. Automation in Construction 30(2013), 126–135.
[2] EI-Diraby, T.A., Lima, C., Feis, B. Domain taxonomy for construction concepts: toward a formal ontology for construction knowledge. Journal of Computing in Civil Engineering 19(2005), 394–406.
[3] Wang, H.-H., Boukamp, F., Elghamrawy, T. Ontology-based approach to context representation and reasoning for managing context-sensitive construction information. Journal of Computing in Civil Engineering 25(2011), 331–346.
[4] EI-Diraby, T.E. Domain ontology for construction knowledge. Journal of Construction Engineering and Management 139(2013), 768–784.
[5] Nepal, M.P., Staub-French, S., Pottinger, R., Webster, A. Querying a building information model for construction-specific spatial information. Advanced Engineering Informatics, EG-ICE 2011 + SI: Modern Concurrent Engineering 26(2012), 904–923.
[6] Zhang, L., Issa, R.R.A. Ontology-based partial building information model extraction. Journal of Computing in Civil Engineering 27(2013), 576–584.
[7] Eastman, C., Lee, J., Jeong, Y., Lee, J. Automatic rule-based checking of building designs. Automation in Construction 18(2009), 1011–1033.
[8] Pauwels, P., Van Deursen, D., Verstraeten, R., De Roo, J., De Meyer, R., Van de Walle, R., Van Campenhout, J. A semantic rule checking environment for building performance checking. Automation in Construction 20(2011), 506–518.
[9] Staub-French, S., Fischer, M., Kunz, J., Paulson, B. An ontology for relating features with activities to calculate costs. Journal of Computing in Civil Engineering 17(2003), 243–254.
[10] Lee, S.-K., Kim, K.-R., Yu, J.-H. BIM and ontology-based approach for building cost estimation. Automation in Construction 41(2014), 96–105.
[11] Ma, Z., Wei, Z. Framework for automatic construction cost estimation based on BIM and ontology technology, in: Proceedings of the CIB W78, 2012.
[12] Ma, Z., Wei, Z., Liu, Z. Ontology-based computerized representation of specifications for construction cost estimation, in: Proceedings of the CIB W78, 2013.
[13] Barforush, A.A., Rahnama, A. Ontology learning: revisited. Journal of Web Engineering 11(2012), 269–289.
[14] Wong, W., Liu, W., Bennamoun, M. Ontology learning from text: a look back and into the future. ACM Comput. Surv. 44(2012), 20:1–20:36.
[15] Sang, A. J. Chinese ontology learning technology based on Text2Onto. Ocean University of China, Shandong, China, 2009.
[16] Wang, D. The Research of Chinese Ontology Learning Based on Web Mining. Taiyuan Universitiy of Technology, Shanxi, China, 2007.
[17] Hou, L. X, Hong, Z. S, He, H. T., Zhao, H & Han. D. Concept lattice reduction application in field of Chinese domain ontology learning. Journal of Jilin University, vol.31, no.6, 2013, pp.621-626.
[18] Zhang, H.-P., Yu, H.-K., Xiong, D.-Y., Liu, Q. HHMM-based Chinese lexical analyzer ICTCLAS, in: Proceedings of the Second SIGHAN Workshop on Chinese Language Processing - Volume 17 (2003), SIGHAN '03. Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 184–187.
[19] Princeton University. About WordNet. 2010. http://wordnet.princeton.edu.
[20] LOPE Lab. About Chinese WordNet, 2011, http://lope.linguistics.ntu.edu.tw/cwn/
[21] Stanford Center for Biomedical Informatics Research. Protégé, http://protege.stanford.edu/
[22] Ma, Z., Liu, Z. BIM-based intelligent acquisition of construction information for cost estimation of building projects. In: Procedia Engineering 85 (2014): 358-367.